Automatic Crowd Analysis from Airborne Images

Beril Sirmacek, Peter Reinartz

German Aerospace Center (DLR), Remote Sensing Technology Institute
Wessling, Germany
Email: Beril.Sirmacek@dlr.de

***Abstract*—Recently automatic detection of people and crowded areas from images became a very important research field, since it can provide crucial information especially for police departments and crisis management teams. Detection of crowd and measuring the density of people can prevent possible accidents or unpleasant conditions to appear. Understanding behavioral dynamics of large people groups can also help to estimate future states of underground passages, shopping center like public entrances, or streets which can also affect the traffic.**

In order to bring a solution to this problem, herein we propose a novel approach using airborne images. Although their resolutions are not enough to see each person in detail, we can still notice a change of color components in the place where a person exists. Therefore, we propose a color feature detection based probabilistic framework. First, we extract local features from invariant chroma bands of the image. Extracted local features behave as observations of the probability density function (pdf) of the crowd to be estimated. Using an adaptive kernel density estimation method, we estimate the corresponding pdf. The estimated pdf gives information about crowded regions, and also helps to extract quantitative measures about them. Our experimental results show that the proposed approach can provide crucial information to police departments and crisis management teams to achieve more detailed observations of crowds to prevent possible accidents or unpleasant conditions in robust and fast manner.

I. Introduction

The quantitative analysis of man events in real time can be of crucial importance to avoid very dense crowds or overload of people in certain areas. Also in some cases it might be of value how many people exist at an event. Understanding behavioral dynamics of large people groups can also help to estimate future states of underground passages, shopping center like public entrances, or streets. Also the effect on vehicle traffic is an important research field.

Due to the importance of the topic, many researchers tried to solve crowd detection problem using street, or indoor cameras which are known as close-range cameras. The early studies in this field were developed from closed-circuit television images [2], [10], [11]. Unfortunately, these cameras can only monitor a few square meters in indoor regions, and it is not possible to adapt those algorithms to street or airborne cameras since the human face and body contours will not appear as clearly as in close-range indoor camera images due to the resolution and scale differences. In order to be able to monitor bigger events researchers tried to develop algorithms which can work on outdoor camera images or video streams. Arandjelovic [1] developed a local interest point extraction based crowd detection method to classify single terrestrial images as crowd and non-crowd regions. They observed that dense crowds produce a high number of interest points. Therefore, they used density of SIFT features for classification. After generating crowd and non-crowd training sets, they used SVM based classification to detect crowds. They obtained scale invariant and good results in terrestrial images. Unfortunately, these images do not enable monitoring large events, and different crowd samples should be detected before hand to train the classifier. Ge and Collins [4] proposed a Bayesian marked point process to detect and count people in single images. They used football match images, and also street camera images for testing their algorithm. It requires clear detection of body boundaries, which is not possible in airborne images. In another study, Ge and Collins [5] used multiple close-range images which are taken at the same time from different viewing angles. They used three-dimensional heights of the objects to detect people on streets. Unfortunately, it is not always possible to obtain these multi-view close-range images for the street where an event occurs. Chao et al. [7] wanted to obtain quantitative measures about crowds using single images. They used Haar wavelet transform to detect head-like contours, then using SVM they classified detected contours as head or non-head regions. They provided quantitative measures about number of people in crowd and sizes of crowd. Although results are promising, this method requires clear detection of human head contours and a training of the classifier. Unfortunately, street cameras also have a limited coverage area to monitor large outdoor events. In addition to that, in most of the cases, it is not possible to obtain close-range street images or video streams in the place where an event occurs. Therefore, to monitor crowded regions in very big outdoor events, the best way is to use airborne images which began to give more information to researchers with the development of sensor technology. Since most of the previous approaches in this field needed clear detection of face or body features, curves, or boundaries to identify crowds which is not possible in airborne images, new approaches are needed to extract crowd information from these images. In a previous study Hinz et al. [6] used registered airborne image sequences to estimate density and motion of people in crowded regions, but their approach did not provide quantitative measures about crowds.

Herein we propose a novel approach to detect people and crowded areas automatically from airborne images. In Fig. 1.(a), we give an example airborne image ($Image_1$ in our

data set) which is taken around a stadium before a soccer match. Unfortunately, airborne image resolutions do not enable to see each single person with sharp details. In Fig. 1.(b), we represent a subpart of the $Image_1$ which is zoomed into a crowded region. As can be seen in this example, it is not possible to find head or body features due to the resolutions and looking angles of airborne images. However, we can still notice a change of color components in the place where a person exists. Therefore, herein we propose a color feature detection based probabilistic framework to detect people, crowded regions and their properties.

(a) (b)

Fig. 1. (a) An airborne image ($Image_1$) including crowded and sparse people groups, (b) Closer view of a crowded region in $Image_1$.

II. Proposed Automatic Crowd Detection and Density Estimation System

We introduce the proposed system in two steps. We begin representing the system with introducing local feature extraction. Then in the second step, we introduce our probabilistic framework to detect crowds and their properties.

A. Local Feature Extraction

Our crowd detection and density estimation method depends on local features extracted from the test image. Due to the low resolutions of airborne images and low scale of people in the scene, structural features (like edges, curves, etc.) cannot help to obtain information about each single person in the image. However, we can still notice a change of color components in the place where a person exists. Therefore, we use color bands of the image for feature extraction.

For local feature extraction, we use features from accelerated segment test (FAST). FAST feature extraction method is especially developed for corner detection purposes by Rosten et al. [12], however it also gives high responses on small regions which are significantly different than surrounding pixels. The method depends on wedge-model-style corner detection and machine learning techniques. For each feature candidate pixel, its 16 neighbors are checked. If there exist nine contiguous pixels passing a set of pixels, the candidate pixel is labeled as a feature location. In FAST method, these tests are done using machine learning techniques to speed up the operation.

For FAST feature extraction from invariant color bands of the image, we first start with converting our RGB test image into CIELab color space. In many computer applications, the CIELab color space is used since it mimics the human visual system [3]. CIELab color space bands are able to enhance different colors best and minimize color variances. After transforming the RGB color image into CIELab color space, again we obtain three bands as L, a, and b [9]. Here, L band corresponds to intensity of the image pixels. a, b bands contain chroma features of the image. These two bands give information about the color information independent of illumination. For illumination invariance, in this study we use only a and b chroma bands of image for local feature extraction. To detect small regions which have significantly different color values than their surroundings, we extract FAST features from a and b chroma bands of the image.

After applying FAST feature extraction method to both a and b chroma bands, we assume (x_i, y_i) $i \in [1, 2, ..., K_i]$ as spatial coordinates of each feature obtained from a and b. Here, K_i is the total number of detected FAST-based features from both two chroma bands. We represent locations of detected local features for our $Image_1$ test image in Fig. 2.(a). Extracted FAST features behave as observations of the probability density function (pdf) of the crowd to be estimated. In the next step, we introduce an adaptive kernel density estimation method, to estimate corresponding crowd pdf.

B. Adaptive Kernel Density Estimation for Crowd Detection

Since we have no pre-information about the street, building, green area boundaries and crowd locations in the image, we formulate the crowd detection method using a probabilistic framework. Assume that (x_i, y_i) is the ith FAST feature where $i \in [1, 2, ..., K_i]$. Each FAST feature indicates a local color change which might be a human to be detected. Therefore, we assume each FAST feature as an observation of a crowd pdf. For crowded regions, we assume that more local features should come together. Therefore knowing the pdf will lead to detection of crowds. For pdf estimation, we benefit from a kernel based density estimation method as Sirmacek and Unsalan represented for local feature based building detection [14].

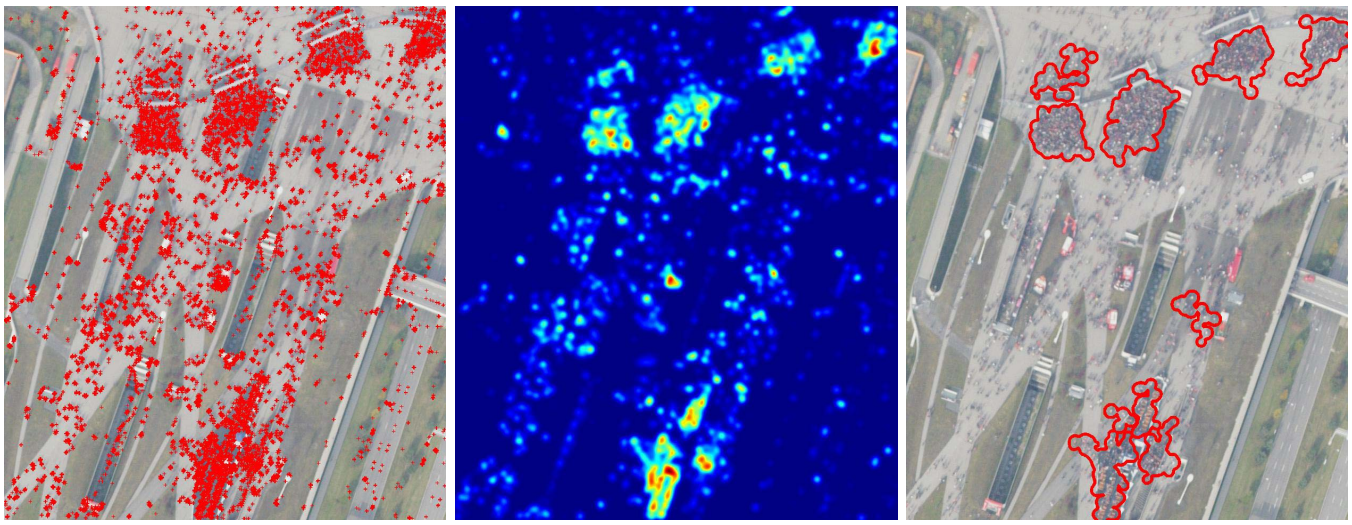
Silverman [13] defined the kernel density estimator for a discrete and bivariate pdf as follows. The bivariate kernel function $[N(x, y)]$ should satisfy the conditions given below;

$$\sum_x \sum_y N(x, y) = 1 \tag{1}$$

$$N(x, y) \geq 0, \forall (x, y) \tag{2}$$

The pdf estimator with kernel $N(x, y)$ is defined by,

$$p(x, y) = \frac{1}{nh} \sum_{i=1}^{n} N(\frac{x - x_i}{h}, \frac{y - y_i}{h}) \tag{3}$$



(a) (b) (c)

Fig. 2. (a) Detected FAST feature locations (from both a and b chroma bands) are represented with red crosses, (b) Estimated probability density function (color coded) generated using FAST feature locations as observations, (c) Automatically detected dense crowds.

where h is the width of window which is also called smoothing parameter. In this equation, (x_i, y_i) for $i = 1, 2, ..., n$ are observations from pdf that we want to estimate. We take $N(x, y)$ as a Gaussian symmetric pdf, which is used in most density estimation applications. Then, the estimated pdf is formed as below;

$$p(x,y) = \frac{1}{R}\sum_{i=1}^{K_i} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma}) \quad (4)$$

where σ is the bandwidth of Gaussian kernel (also called smoothing parameter), and R is the normalizing constant. In Fig. 2.(b), we show estimated pdf function for our sample image for a constant σ value ($\sigma = \sqrt{500}$). Represented pdf function is color coded, which means yellow-red regions show high probability values, and dark blue regions show low probability values. As can be seen in this figure, crowded areas have very high probability values, and they are highlighted in estimated pdf. We use Otsu's automatic thresholding method on this pdf to detect regions having high probability values [8]. After thresholding our pdf function, in obtained binary image we eliminate regions with an area smaller than 1000 pixels since they cannot indicate large human crowds as we aimed in this study. The resulting binary image $B_c(x, y)$ holds dense crowd regions. For $Image_1$, boundaries of detected crowd regions are represented on original input image in Fig. 2.(c).

In kernel based density estimation the main problem is how to choose the bandwidth of Gaussian kernel for a given test image, since the estimated pdf directly depends on this value. For instance, if the resolution of the camera increases or if the altitude of the plane decreases, pixel distance between two persons will increase. That means, Gaussian kernels with larger bandwidths will make these two persons connected and will lead to detect them as a group. Otherwise, there will be many separate peaks on pdf, but we will not be able to find large hills which indicate crowds. As a result, using a Gaussian kernel with fixed bandwidth will lead to poor estimates. Therefore, bandwidth of Gaussian kernel should be adapted for any given input image.

In probability theory, there are several methods to estimate the bandwidth of kernel functions for given observations. One well-known approach is using statistical classification. This method is based on computing the pdf using different bandwidth parameters and then comparing them. Unfortunately, in our field such a framework can be very time consuming for large input images. The other well-known approach is called balloon estimators. This method checks k-nearest neighborhoods of each observation point to understand the density in that area. If the density is high, bandwidth is reduced proportional to the detected density measure. This method is generally used for variable kernel density estimation, where a different kernel bandwidth is used for each observation point. However, in our study we need to compute one fixed kernel bandwidth to use at all observation points. To this end, we follow an approach which is slightly different from balloon estimators. First, we pick $K_i/2$ number of random observations (FAST feature locations) to reduce the computation time. For each observation location, we compute the distance to the nearest neighbour observation point. Then, the mean of all distances give us a number l (calculated 105.6 for $Image_1$). We assume that variance of Gaussian kernel (σ^2) should be equal or greater than l. In order to guarantee to intersect kernels of two close observations, we assume variance of Gaussian kernel as $5l$ in our study. Consequently, bandwidth

of Gaussian kernel is estimated as $\sigma = \sqrt{5l}$. For a given input image, that value is computed only one time. Then, the same σ value is used for all observations which are extracted from images of the same camera and the same flight altitude. The introduced automatic kernel bandwidth estimation method, makes the algorithm robust to scale and resolution changes.

After detecting dense crowds automatically, next we want to extract quantitative measures from detected crowds for more detailed analysis. First, we start with analyzing detected dense crowds. Since they indicate local color changes, detected features can give information about number of people in crowded areas. Unfortunately, number of features in a crowd region do not give the number of people directly. In most cases, shadows of people or small gaps between people also generate a feature. Besides, two neighbour features might come from two different chroma bands for the same person. In order to decrease counting errors coming from these features, we follow a different strategy to estimate the number of people in detected crowds. We use a binary mask $B_f(x, y)$ where feature locations have value 1. Then, we dilate $B_f(x, y)$ using a disk shape structuring element with a radius of 2 to connect close feature locations. Finally, we apply connected component analysis to mask, and we assume the total number connected components which are laying in a crowd area as the number of people (N). In this process, slight change of radius of structuring element does not make a significant change in estimated people number N. However, an appreciable increase in radius can connect features coming from different persons, and that decreases N which leads to poor number of people estimates.

If the resolution of the input image is known, using estimated number of people in crowd, the density of people (d) can also be calculated. Lets assume, $B_c^j(x, y)$ is the jth connected component in $B_c(x, y)$ crowd mask. We calculate crowd density for jth crowd as $d^j = N/(\sum_X \sum_Y B_c^j(x, y) \times a)$, where X and Y are the numbers of pixels in the image in horizontal and vertical directions respectively, and a is the area of one pixel as m^2.

III. Experiments

To test our method, we use two different open-air concert images, a Munich beer festival image, and a stadium entrance data set which includes 43 images taken in different times. However we have obtained successful detection results for whole stadium entrance image data set, we provide detection result only for $Image_1$ in Fig. 2.(c) due to the page limitations. To obtain a measure about the performance of the algorithm, we have generated groundtruth data for four dense crowds in $Image_1$ which are represented in Fig. 3. Since even for human observer it is hard to count the exact number of people in crowds, we have assumed mean of counts of three human observers as groundtruth. In Table I, we compare automatically detected number of people (N), and density (d) with groundtruth data (N_{gth} and d_{gth} respectively) for each crowd. Similarity of our measures with groundtruth shows the high performance of the proposed approach.

In Fig. 4, we also represent three sample results using our two open-air concert images and a Munich beer festival image. Obtained results indicate robustness of algorithm in detection of dense crowds even in very different environments.

Fig. 3. Labels of detected crowds which are used for performance analysis.

TABLE I
COMPARISON OF GROUNDTRUTH AND AUTOMATICALLY DETECTED PEOPLE NUMBER AND DENSITY ESTIMATION RESULTS FOR TEST REGIONS IN $Image_1$.

	$REGION_1$	$REGION_2$	$REGION_3$	$REGION_4$
N	139	211	115	102
N_{gth}	132	180	114	98
d	0.81	0.74	0.68	0.76
d_{gth}	0.76	0.63	0.67	0.73

IV. Conclusions and Future Work

In order to solve crowd detection and analysis problem, herein we propose a novel approach to detect crowded areas automatically from airborne images. Although resolutions of airborne images are not enough to see each person with sharp details, we can still notice a change of color components in the place where a person exists. Therefore, we benefit from local features which are extracted from illumination invariant chroma bands of the image. Assuming extracted local features as observation points, we generated a probability density function using Gaussian kernel functions with constant bandwidths. For deciding bandwidth of Gaussian kernel to be used, we introduced an adaptive method. In this way, we obtained a robust algorithm which can cope with input images having different resolutions. By automatically thresholding obtained pdf function, dense crowds are robustly detected. After that, feature locations at detected crowds are analyzed to estimate number of people and people density in crowded regions. We have tested our algorithm on a large stadium entrance image data set, two different open-air concert images, and a Munich beer festival image. Our experimental results indicate possible usage of the algorithm in real-life events. We believe that, the proposed study has very high importance since it provides real-time quantitative measures about crowds automatically.

References

[1] O. Arandjelovic, "Crowd detection from still images," *British Machine Vision Conference (BMVC'08)*, Sep. 2008.

[2] A. Davies, J. Yin, and S. Velastin, "Crowd monitoring using image processing," *IEEE Electronic and Communications Engineering Journal*, vol. 7 (1), pp. 37–47, 1995.

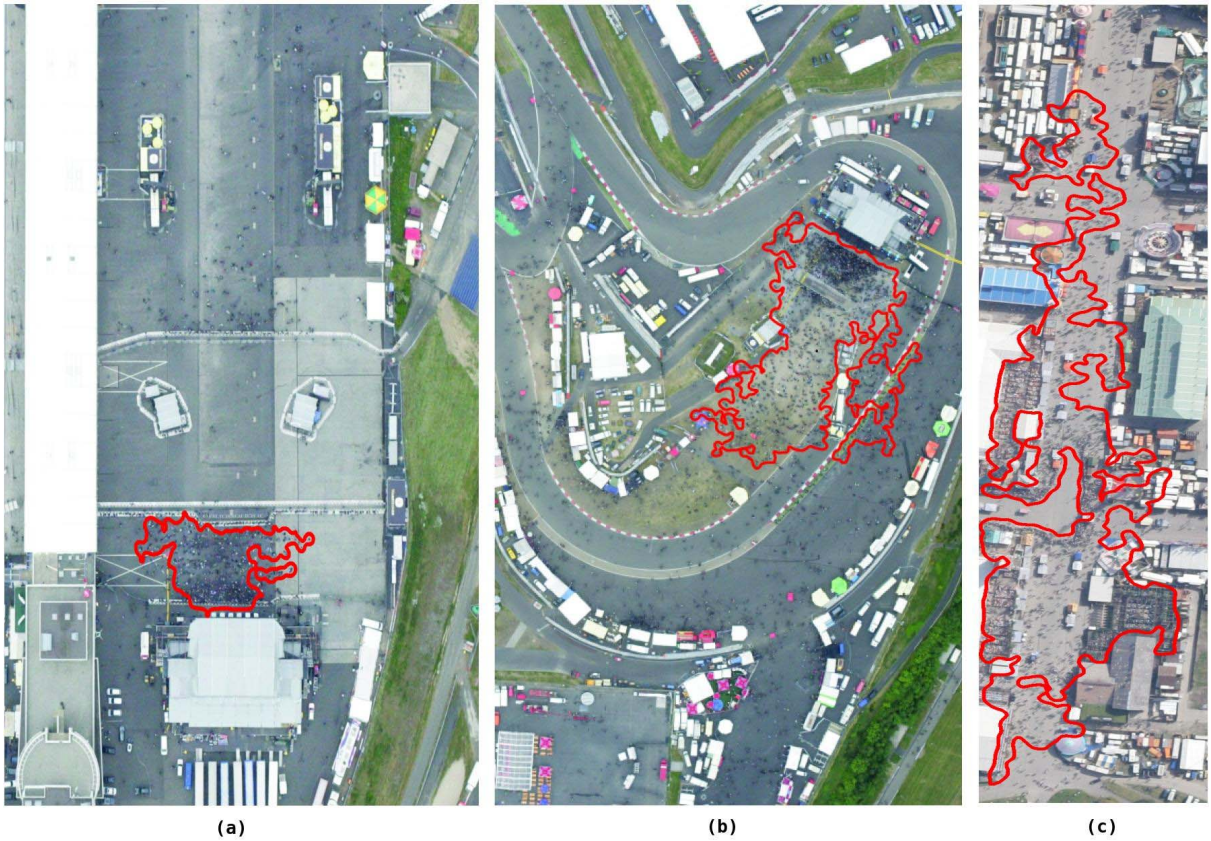


Fig. 4. (a) Crowd detection result for the first concert image, (b) Crowd detection result for the second concert image, (c) Crowd detection result for Munich beer festival image.

[3] M. Fairchild, "Color appearance models," *Addison-Wesley*, 1998.

[4] W. Ge and R. Collins, "Marked point process for crowd counting," *IEEE Computer Vision and Pattern Recognition Conference (CVPR'09)*, pp. 2913–2920, 2009.

[5] ——, "Crowd detection with a multiview sampler," *European Conference on Computer Vision (ECCV'10)*, 2010.

[6] S. Hinz, "Density and motion estimation of people in crowded environments based on aerial image sequences," *ISPRS Hannover Workshop on High-Resolution Earth Imaging for Geospatial Information*, 2009.

[7] S. Lin, J. Chen, and H. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 31 (6), pp. 645–654, Nov. 2001.

[8] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on System, Man, and Cybernetics*, vol. 9 (1), pp. 62–66, 2009.

[9] G. Paschos, "Perceptually uniform color spaces for color texture analysis: an empirical evaluation," *IEEE Transactions on Image Processing*, vol. 10, pp. 932–937, 2001.

[10] C. Regazzoni and A. Tesei, "Local density evaluation and tracking of multiple objects from complex image sequences," *Proceedings of 20th International Conference on Industrial Electronics, Control and Instrumentation (IECON)*, vol. 2, pp. 744–748, 1994.

[11] ——, "Distributed data fusion for real time crowding estimation," *Signal Processing*, vol. 53, pp. 47–63, 1996.

[12] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Learning*, vol. 32 (1), pp. 105–119, Nov. 2010.

[13] B. Silverman, "Density estimation for statistics and data analysis," *1st Edition*, vol. London, UK, Chapman and Hall, 1986.

[14] B. Sirmacek and C. Unsalan, "A probabilistic framework to detect buildings in aerial and satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, 2010.